

AI-generated News May Be Coming in Your Language: A Case Study of Italian

Anonymous ACL submission

Abstract

The number of available Large Language Models (LLMs) is growing steadily for English, but less so for other languages. This may create the impression that the potentially harmful applications of LLMs are also limited to English. We present a case study of Italian to investigate the possibility of generating fluent news-like texts with Llama, an existing LLM that was mostly trained on English, with only 40K Italian news articles for fine-tuning. We find that this is sufficient for producing texts that native speakers of Italian struggle to identify as synthetic. We also experiment with two statistical methods of detecting synthetic texts (log-likelihood and DetectGPT), finding that they perform better than human raters. However, these methods are unusable in practice, since they require access to token likelihood information.

1 Introduction

The modern Large Language Models (LLMs) can generate increasingly fluent and plausible-sounding texts, which sparks concerns about their potential misuse by bad actors. One of the emerging problems is “content farms” (Hurst, 2023): news-like sites filled with synthetic texts that are not necessarily serving some specific misinformation campaign, but are plausible looking enough to generate web traffic and benefit from displaying ads – while not providing people with authentic news content. But since high-quality LLMs are currently available primarily for English, it may seem that other languages are safe from such content for now.

We present a case study of Italian, in which we fine-tune Llama (Touvron et al., 2023), a general-purpose LLM that was mostly trained on English, on only 40k Italian news texts. We then try to detect the synthetic Italian text with statistical methods (Section 4) and by human readers (Section 5), and we find that the accuracy of synthetic news text detection by the native speakers of Italian for

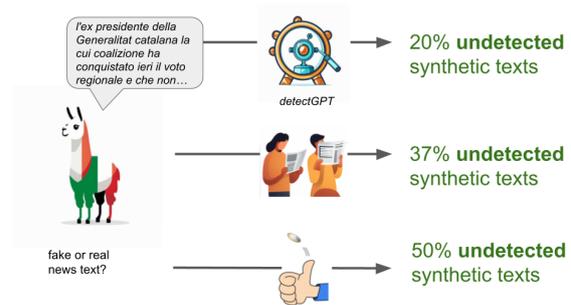


Figure 1: Detecting synthetic Italian news text generated by fine-tuned Llama-65B: error rates for DetectGPT, native speakers of Italian and random guess.

Llama65B is only 64% (vs 50% random guess), as shown in Figure 1. We find that automated methods (DetectGPT, log-likelihood) perform better, but they would be unusable in practice, since they require access to token likelihood information. This is worrying, and we hope that our findings would spur similar investigations for other languages, and highlight the urgency of developing model-agnostic methods for synthetic text detection.

2 Experimental setup

Model fine-tuning. Our goal is to test the feasibility of generating fluent news-like text in Italian by fine-tuning an English LLM, rather than investing into a high-quality monolingual LLM for Italian. Presumably such work would start with a openly accessible high-performing model. We choose the Llama model (Touvron et al., 2023), in 7B and 65B parameter versions.

Our choice of Italian is motivated by the possibility of qualitative analysis in this language by the authors. It is not a low-resource language, but there are few Italian LLMs, and even the largest of them (Sarti and Nissim, 2022) is much smaller than the current English LLMs. We know that Llama was exposed to at least Italian Wikipedia (Touvron

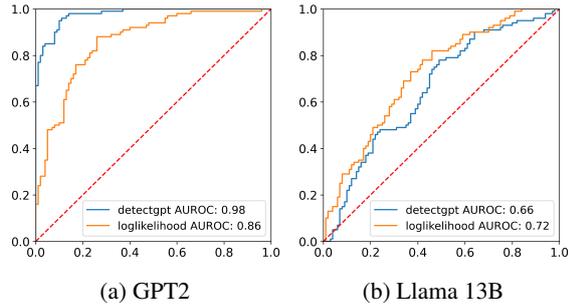


Figure 2: ROC curve for DetectGPT and log-likelihood. In (a) for the GPT2 model over 100 samples from xsum (coherent with (Mitchell et al., 2023)), in (b) for llama13b model over 100 samples from xsum.

et al., 2023, p.2), and that the other sources of Llama pre-training text, such as C4 corpus (Raffel et al., 2020), deliberately attempted to filter out non-English text. Hence the main source of Italian in Llama was likely Wikipedia.

We fine-tune both Llama models on a randomly chosen 40K subset of the CHANGE-it news dataset (Mattei et al., 2020). The articles are arranged in training sequences composed of 128 tokens, adding subsequent segments one after the other.

The training was performed on 8 nodes, each with 4 V100 GPUs with 16GB VRAM. The effective batch size is 128, real batch size 2, 64 accumulation steps. The maximum learning rate is 0.0005, with a one-cycle scheduler without warmup.

Due to memory constraints, instead of the most effective AdamW optimizer we use simple Stochastic Gradient Descent. We train for 60,000 steps (120,000 samples split into 3 epochs of 40,000) saving checkpoints every 10,000 steps (20,000 samples). This takes approximately 3 days.¹

Methodology for automated detection of synthetic text. We experiment with two approaches for the detection of synthetic text: the measurement of the *log-likelihood* of a text passage according to the model itself, and the *DetectGPT* score proposed by Mitchell et al. (2023). Both these approaches assume that one has access to the likelihood of each token composing a sentence according to the model whose “authorship” is under analysis. Similarly to Jawahar et al. (2020); Sadasivan et al. (2023); Chakraborty et al. (2023), we attempt the zero-shot

¹ The actual run lasted 5 days as we wasted approximately 2 days due to exploding loss. Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO₂eq/kWh. A cumulative of 2,304 hours of computation was performed, and total emissions are estimated to be 298.6 kgCO₂eq (Luccioni et al., 2019).

	<i>DetectGPT</i>			<i>log-likelihood</i>
	t5-base	t5-3b	t5-11b	
Llama 13b	66%	71%	75%	78%
Llama 65b	-	62%	66%	70%

Table 1: AUROC achieved by *DetectGPT* (varying bootstrapping model) and *log-likelihood* on the xsum dataset for Llama 13B and Llama 65B.

detection of artificially generated text.

The core idea behind the DetectGPT score is to normalize the likelihood of a sentence on the basis of modifications of that same sentence, that were generated by a different model (which we refer to as the *bootstrap model*). In our case, we are interested in the likelihood of a synthetic sentence estimated by Llama, vs the normalized likelihood that Llama assigns to modifications of that same sentence that we generate with t5 as the bootstrap model (Raffel et al., 2020). We refer the reader to Mitchell et al. (2023) for details of this method.

To measure the ROC of both *DetectGPT* and *log-likelihood* we always select 100 random sentences and generate² another 100 by selecting the first 30 tokens of each sentence and generating a new one using this chunk as a prompt for the model under analysis. After this, we clip all sentences to 150 tokens³, measure the score for each sample, and compute the AUROC on all the 200 sentences (half human written and half machine-completed).

To compare the detection approaches for the original and fine-tuned Llama, we use the CHANGE-it test dataset for Italian (Mattei et al., 2020), and the xsum dataset for English (Narayan et al., 2018). As the bootstrap model, we use *it5-large* (Sarti and Nissim, 2022) for Italian, and different versions of *t5* (Raffel et al., 2020) for English.

3 Synthetic text detection in English

We start by replicating a core result by Mitchell et al. (2023): on xsum data, DetectGPT outperforms the *log-likelihood* in detecting GPT2 text (Radford et al., 2019) (see Figure 2a).

We apply the same methodology to sentences obtained using Llama 13B and Llama 65B. For Llama, we were unable to get *DetectGPT* to achieve a higher AUROC than the *log-likelihood*. We believe this to be due to the stronger performance of Llama

²All generation tasks are performed using Nucleus Sampling for decoding (Holtzman et al., 2020).

³Due to different tokenizers, this step results in sentences with a varying number of words, but a similar length.

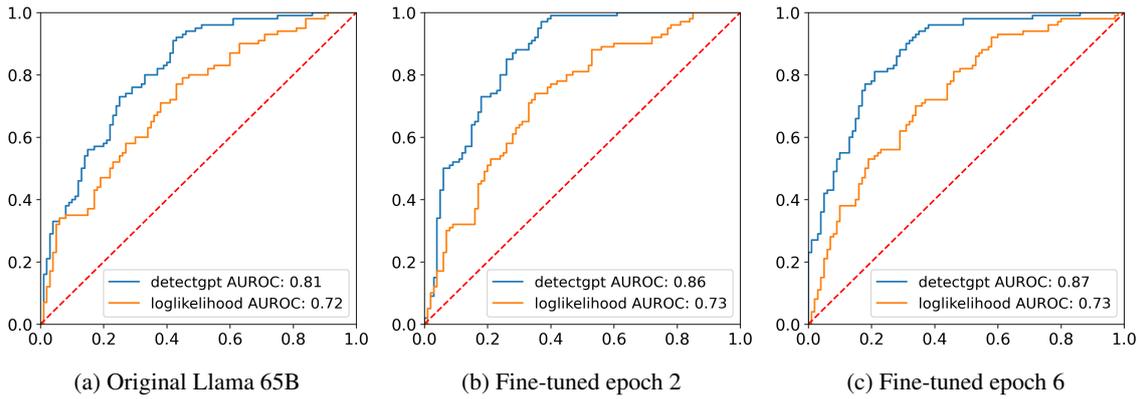


Figure 3: ROC curve for *DetectGPT* and *log-likelihood*. In (a) for Llama 65B measured over 100 sentences from the CHANGE-it data-set (Italian), in (b) the same measure for Llama 65B model after 20,000 fine tuning steps on CHANGE-it training set and in (c) after 60,000 fine-tuning steps.

compared to the t5 model used to generate new sentences (Figure 2b). This suggests that English text generated by Llama is harder to detect.

To test this hypothesis we measure the importance of the bootstrap model in this case. Table 1 shows the AUROC of *DetectGPT* depending on the bootstrap model, a larger t5 model leads to higher AUROC and it is therefore possible that a stronger one is needed.

We repeat the experiment with both *DetectGPT* and *log-likelihood* at various temperature settings (0.6, 0.8, 1.0), and we find a strong sensitivity to this hyper-parameter, which merits further investigation. The numerical results for our experiments can be found in App. A.

4 Detection of Synthetic Italian News

We experiment with three models: the original Llama baseline (65B pre-trained model with no extra training), and two versions of our Llama fine-tuned on Italian news, after 20k and 60k steps.

Figure 3 shows the AUROC of *DetectGPT* and *log-likelihood* for all our models. All of them show very close AUROC for *log-likelihood* while the two fine-tuned ones show higher values for *DetectGPT*. This is in contrast with the intuition that fine-tuned models should generate text that is more plausible in the context where they are adapted.

However, by qualitative analysis we found one pattern that allows to easily detect both the original 7B and the 65B models: they unexpectedly switch to English, whereas the fine-tuned models do not. Here is an example (English translation here⁴):

⁴"[...]the former president of Catalan Generalitat, whose coalition won the regional election yesterday, and who

"[...]l'ex presidente della Generalitat catalana la cui coalizione ha conquistato ieri il voto regionale e che non"

- **Pre-trained:** *vuole rinunciare alla secessione. In the 6-week period prior to 12/06/19"*
- **Fine-tuned:** *aveva perso tempo per dire la sua. Da Bruxelles, dove si trova da allora"*

See App. C for more examples. These sentences, easily found by a native speaker of Italian, are still missed by the detection algorithms. Thus, they can be naturally adversarial examples for *DetectGPT* and *log-likelihood* detection.

Although the *DetectGPT* and *log-likelihood* perform relatively well in our tests, we stress that this result indicates a measure of the difficulty of this task, rather than a solution to synthetic news detection. These methods require access to the token likelihood information, which is unavailable for published texts generated by an unknown model. We remark that *DetectGPT* score can be turned into an accuracy measurement by fixing a threshold, for a direct comparison with the human evaluation accuracy. In our case this would give us an accuracy $\approx 80\%$ using the median⁵ score as the threshold for fine-tuned Llama 65B.

5 Detection of Synthetic Italian News by Humans

To assess whether native speakers of Italian would be able to easily identify the synthetic news text,

Pre-trained: *does not want to give up the secession. In the 6-week period prior to 12/06/19"*, **Fine-tuned:** *does not waste time to mention his opinion. From Brussels, where he resides since"*.

⁵We choose this threshold knowing that the dataset is balanced and that *DetectGPT* is monotonic, otherwise we would need to tune it.

Model	Accuracy	STD	Fleiss κ
<i>Llama 7B</i> pretrain	83.2	7.0	36.45
<i>Llama 7B</i> finetune	69.5	12.2	22.30
<i>Llama 65B</i> pretrain	73.7	5.8	33.01
<i>Llama 65B</i> finetune	64.2	11.2	20.56

Table 2: Accuracy achieved by human raters in assessing human-written versus machine generated news. We report the overall accuracy and the standard deviation.

we created 4 different surveys with 100 questions each, split into 5 sessions with 20 questions in each session (van der Lee et al., 2021).⁶

Each version of the survey corresponded to one of the 4 models: Llama 7B and 65B, with and without fine-tuning on Italian news. Each question had the following structure: *"Text B follows text A, do you think text B is written by a machine?"* The raters were asked to assign a value from the given scale: 1 (*certainly human-written*), 2, 3, 4, 5 (*certainly machine-generated*). In 50% of the questions, both text A and B came from an original human-written news article. The texts were randomly selected from the CHANGE-it test dataset. The topics included daily national political (e.g. politicians’ declarations) and general news (e.g. climate catastrophes), and relevant international news (e.g. European leadership meetings).

A binary score was assigned to each question, indicating whether text B is perceived as human-generated or machine-generated. We computed the mean human-score of each question, which reflects the average perception of the raters. If the mean score is higher than 3 (the middle score on the scale), text B is marked as machine-generated. We also experimented with a different thresholding approach that showed similar results (see App. B).

Table 2 shows the outcome of our analysis. Since the raters’ accuracy in detecting news generated by the largest fine-tuned Llama 65B is as low as 64%, we can answer our research question positively: **Llama can be fine-tuned to generate hard-to-detect news-like text in Italian.** We only used 40K samples for fine-tuning, so it is likely that even more plausible synthetic text could be created with more data.

⁶The raters (native speakers of Italian) were anonymously recruited and compensated through Prolific, an online crowdsourcing platform. Each text completion was assessed by 5 raters. The 20-question sessions took 8 mins and 23 secs on average, and the raters were compensated at 7.37 Euro per session.

Overall, the raters’ accuracy exhibits two familiar trends: the smaller models perform worse than the larger ones, and fine-tuned models outperform vanilla Llama. Fine-tuned models also induce higher variation between the human raters, since more of them are deceived. Interestingly, the small 7B version, fine-tuned on Italian, is identified by raters with accuracy close to the larger 65B without fine-tuning.

Table 2 also reports the inter-rater agreement. Fleiss κ (Fleiss et al., 1971) is in the range between 22%-36%, indicating a “moderate” agreement (Landis and Koch, 1977). The raters agree more strongly when they rate non-fine-tuned models. This aligns with our previous observation (Section 4) that such models show certain pitfalls easily spotted by humans, e.g. falling back to English.

We manually inspected 100 generated instances (25 per model), and we found that in that sample, 46/100 examples had no obvious issues with language, but their content was hallucinated. Among the easy-to-detect issues we found 7 examples where the generated text contradicted the prompt, 8 switching to English, 18 containing grammatical errors and 21 contained Italian expressions that were not grammatically wrong, but sounded unnatural. See see App. C for examples for each model.

6 Conclusion

This short paper examines the possibility of generating plausible-sounding news-like text in Italian, a language that does not have a high-quality monolingual LLM, by fine-tuning Llama, a mostly-English LLM. We find that after fine-tuning Llama 65B on only 40K Italian news texts, native speakers of Italian have only $\approx 64\%$ accuracy on synthetic text detection task. The same model without fine-tuning is as hard to detect as the 7B fine-tuned model ($\approx 73\%$ accuracy).

We also show that two statistics-based automatic text detection approaches (based on DetectGPT score and log-likelihood) perform better than human raters ($\approx 80\%$ accuracy with heuristic threshold on DetectGPT score for the largest fine-tuned model), but these methods cannot be used to identify real “content farms”, since they require access to the token likelihood information. Our study highlights the urgency of work on developing reliable, model-agnostic methods of synthetic text detection.

7 Limitations

This short paper presents a case study on a single language. It should not be interpreted as claiming that it is possible to generate plausible-sounding text in any language, by fine-tuning a mostly-English model like Llama. But our results suggest that it *may* be possible, at least for languages with a similar level of coverage in datasets used for training LLMs. More research is needed to establish both the factors impacting the success of such transfer, and better methods to detect synthetic texts.

For Llama specifically, according to Touvron et al. (2023, p.2), the model was exposed to Italian Wikipedia in pre-training. Italian Wikipedia currently has about 500K articles⁷. For other sources included in Llama, such as C4 (Raffel et al., 2020), we cannot exclude the possibility that there was some Italian – but deliberate effort was made to filter out non-English texts, and so we assume that there was at least not much contamination. But other languages in Llama with about the same amount of Wikipedia data as Italian are Polish and Dutch ($\approx 500K$ articles). In the $\approx 400K$ range there are Spanish and Portuguese, and at about $\approx 300K$ – Russian and Swedish. A future study could explore how the amount of Wikipedia data, the amount of fine-tuning data, and typological distance from English impact the success of the transfer.

In scope of this short paper we experimented with two statistical detection methods, but there are others, including supervised detection. This does not invalidate our conclusions and the general answer to our research question, but it could be expanded in the future work.

Our selection of human raters was based solely on Italian as their native language. Future work could investigate whether the results would differ across different occupations and education levels.

Our human evaluation protocol considers the setting where the model is prompted with the first 30 tokens of a real human-written article, because the model is not trained using the articles headlines but just to generate news, to make the adaptation from English to Italian simpler. Another scenario to be tested in future work is when the model is prompted with headlines (authored by the content farm owner or auto-generated). That could affect the quality of the generated text or the ease of its detection.

⁷https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

Finally, our study focuses on the possibility to create plausible-sounding news-like text that could be used by “content farms”, rather than text created for specific misinformation campaigns or to spread conspiracy theories. It is possible that, similarly to human-authored fake news, the human raters would be more likely to doubt the authenticity of the article when it had some big factual claims that were easy to check. This factor also remains to be explored in future work.

8 Broader Impacts

This work aims to highlight a potential problem for the information infrastructure of worldwide communities that may consider themselves safe from plausible-looking synthetic text due to the lack of high-quality monolingual models for their languages. We show that the Llama model, exposed only to Italian Wikipedia and 40K news articles for fine-tuning, is sufficient for generating very plausible-looking synthetic news. We hope that this work would spark similar investigations for other languages, and highlight the urgency of development of reliable and model-agnostic methods for detecting synthetic text.

This work is based on the publicly available models (Radford et al., 2019; Touvron et al., 2023; Raffel et al., 2020) and resources (Narayan et al., 2018; Mattei et al., 2020), and documents its emissions (Footnote 1), annotation procedure and compensation to the human raters (Footnote 6). The code to reproduce our experiments accompanies the submission will be publicly available with the publication of the paper.

References

- Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. [On the possibilities of AI-generated text detection.](#)
- J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation.](#) In *International Conference on Learning Representations.*
- Luke Hurst. 2023. [AI-generated fake news websites driving spread of misinformation.](#)
- Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020. [Automatic detection of machine generated text: A critical survey.](#)

378 J. Richard Landis and Gary G. Koch. 1977. The mea-
379 surement of observer agreement for categorical data.
380 *biometrics*, pages 159–174.

381 Sasha Luccioni, Victor Schmidt, Alexandre Lacoste,
382 and Thomas Dandres. 2019. [Quantifying the carbon](#)
383 [emissions of machine learning](#). In *NeurIPS 2019*
384 *Workshop on Tackling Climate Change with Machine*
385 *Learning*.

386 Lorenzo De Mattei, Michele Cafagna, Felice
387 Dell’Orletta, Malvina Nissim, and Albert Gatt.
388 2020. [CHANGE-IT @ EVALITA 2020: Change](#)
389 [headlines, adapt news, generate \(short paper\)](#). In
390 *Proceedings of the Seventh Evaluation Campaign*
391 *of Natural Language Processing and Speech Tools*
392 *for Italian. Final Workshop (EVALITA 2020), Online*
393 *event, December 17th, 2020*, volume 2765 of *CEUR*
394 *Workshop Proceedings*. CEUR-WS.org.

395 Eric Mitchell, Yoonho Lee, Alexander Khazatsky,
396 Christopher D. Manning, and Chelsea Finn. 2023.
397 [DetectGPT: Zero-shot machine-generated text detec-](#)
398 [tion using probability curvature](#).

399 Shashi Narayan, Shay B. Cohen, and Mirella Lapata.
400 2018. [Don’t give me the details, just the summary!](#)
401 [topic-aware convolutional neural networks for ex-](#)
402 [treme summarization](#). In *Proceedings of the 2018*
403 *Conference on Empirical Methods in Natural Lan-*
404 *guage Processing*, pages 1797–1807, Brussels, Bel-
405 gium. Association for Computational Linguistics.

406 Alec Radford, Jeffrey Wu, Rewon Child, David Luan,
407 Dario Amodei, and Ilya Sutskever. 2019. Language
408 models are unsupervised multitask learners. *Techni-*
409 *cal report*.

410 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine
411 Lee, Sharan Narang, Michael Matena, Yanqi Zhou,
412 Wei Li, and Peter J. Liu. 2020. Exploring the limits
413 of transfer learning with a unified text-to-text trans-
414 former. *J. Mach. Learn. Res.*, 21(1).

415 Vinu Sankar Sadasivan, Aounon Kumar, Sriram Bala-
416 subramanian, Wenxiao Wang, and Soheil Feizi. 2023.
417 [Can AI-generated text be reliably detected?](#)

418 Gabriele Sarti and Malvina Nissim. 2022. [It5: Large-](#)
419 [scale text-to-text pretraining for italian language un-](#)
420 [derstanding and generation](#).

421 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier
422 Martinet, Marie-Anne Lachaux, Timothée Lacroix,
423 Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal
424 Azhar, Aurelien Rodriguez, Armand Joulin, Edouard
425 Grave, and Guillaume Lample. 2023. [Llama: Open](#)
426 [and efficient foundation language models](#).

427 Chris van der Lee, Albert Gatt, Emiel van Miltenburg,
428 and Emiel Kraemer. 2021. [Human evaluation of](#)
429 [automatically generated text: Current trends and best](#)
430 [practice guidelines](#). *Computer Speech & Language*,
431 67:101151.

model	accuracy	std
<i>Llama 7B</i> pretrain	85.3	6.1
<i>Llama 7B</i> finetune	73.7	8.8
<i>Llama 65B</i> pretrain	72.6	3.9
<i>Llama 65B</i> finetune	65.3	16.9

Table 3: Accuracy achieved by human raters in assess-
ing human-written versus machine generated news. We
report the overall accuracy and the standard deviation.

Temperature	<i>DetectGPT</i>	<i>log-likelihood</i>
0.6	48%	86%
0.8	63%	73%
1.0	54%	52%

Table 4: AUROC achieved by *DetectGPT* and *log-*
likelihood on Llama 13B varying the temperature used
while generating the synthetic sentences.

A Temperature Ablation

To establish a fair comparison with *DetectGPT*
while testing Llama we perform an ablation study
based on varying the temperature used in gener-
ation. [Table 4](#) shows different AUROC values
for different temperatures it appears that there is a
strong sensitivity of the detection methodologies
to this hyper-parameter and we plan to investigate
it further. The value 0.8 where *DetectGPT* and
log-likelihood are more aligned, is also the value
reported in the Llama repository.

B Further Human Evaluation Metric

To compute the readers accuracy on identifying
machine generated texts, we threshold the average
score assigned to a sample to obtain a binary la-
bel. While in [Section 5](#) we show the results using
3 as a threshold, the mean possible rating. We
show the same results using a threshold the average
score in the survey. We compute the **scaled mean**:
the mean score of all questions in a 100-question
survey, used as the threshold for distinguishing
between human-generated and machine-generated
texts. [Table 3](#) shows ther results which are well
aligned with [Table 2](#).

C Generation examples

Models Switching to English

Original Text: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre unici superstiti dell'ultimo naufragio avvenuto a 50 chilometri a nord est dalla Libia, costato la vita a 117 persone, che erano partite giovedì a bordo di un gommone sgonfiatosi a circa undici ore di navigazione. I passeggeri sono morti annegati, mentre I tre naufraghi sono riusciti a trovare un appiglio fino all'arrivo dei soccorsi italiani. I tre stranieri...

Pre-trained Llama: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre naufraghi di Tortuga, recuperati da un mercantile libanese. Non è ancora stata resa nota l'identità nazionale dei naufraghi, ma non mancano inchieste di persone che si sono rivolte ai carabinieri per cercare i loro cari. Lampedusa, 38 migrants - mostly from Tunisia - arrived safely on the island ...

Fine tuned epoch 6 Llama: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre bambini rimasti uccisi durante lo stesso sbarco, due di questi non appartenevano a una famiglia, ma erano stati imbarcati separati dai genitori per il viaggio da Libia verso l'Italia. È stato appunto il maggiore Damiano Maiorino, di stanza a Lampedusa, a scoprire i morti. "Non sono morti cadendo in mare da una barca troppo affollata. Hanno subito commentato ...

Original Text: L' Onu è pronta a fare la sua parte alla conclusione del vertice europeo, ma il segretariato non ha in agenda alcuna azione militare, anche se ci possono essere alcune similarità con l'operazione in Somalia ". In una intervista a Qn il vicesegretario generale dell'Onu Jan Eliasson apre "all'uso della forza" per fermare gli sbarchi dei clandestini ed evitare nuove stragi nel Mediterraneo. Ma, al tempo stesso, sottolinea "una differenza qualitativa" con l'operazione in Somalia: "L'attività dei pirati era solo criminale, mentre in Libia abbiamo un urgentissimo problema umanitario prima che criminale". Parlando dell'affondamento preventivo dei barconi nei porti libici...

Pre-trained Llama: L' Onu è pronta a fare la sua parte alla conclusione del vertice europeo, ma il segretariato non ha in agenda alcuna azione militare, anche se ci possono essere alcune iniziative sul campo degli aiuti umanitari", ha affermato un portavoce dell'organizzazione internazionale. Per il segretario generale, "l'Italia deve avere una posizione di leader ed essere la prima a proporre una risposta alla crisi europea, visto anche che l'Italia ha avuto un ruolo molto importante nei negoziati sul tema del contingente europeo". UNITED NATIONS, January 31 – Without waiting for the European Union's decision on...

Fine tuned epoch 6 Llama: L' Onu è pronta a fare la sua parte alla conclusione del vertice europeo, ma il segretariato non ha in agenda alcuna azione militare, anche se ci possono essere alcune azioni umanitarie". Così parla il portavoce dell'organizzazione internazionale, Stephane Dujarric, che aggiunge: "Un vertice europeo è una questione interna". La conferma che l'Onu non è pronta a seguito di una intervista rilasciata dal segretario generale aggiunto, Jeffrey Feltman, a radio 24. "Non abbiamo notizie di alcuna iniziativa che possa condurre all'attuazione di un mandato militare contro l'islamic State sul territorio siriano, anche se ritengo che si stia..."

Table 5: Real and generated sentences in Italian from Llama 65B pretrained and Llama 65B fine-tuned on Italian all starting from the same **prompt**. The sentences generated by Llama 65B pretrained are not detected by either DetectGPT or log-likelihood, despite being partially written in English – which could not happen in a real Italian news article.

Llama 7B Pre-Trained

Generated Text: I dodici ragazzi rimasti intrappolati con il loro allenatore nella grotta di Tham Luang, nella provincia di Chiang Rai, in Thailandia, per oltre due settimane, hanno lasciato oggi l'ospedale dove erano ricoverati. La polizia ha riferito che i ragazzi sono stati sottoposti a un test del DNA che ha dimostrato che tutti erano discendenti di due genitori, ma non vi è nessuna prova che uno o più ragazzi non sia rimasto tra loro durante la fuga dalla grotta. Il...

Translation: The 12 boys who were trapped with their coach in the Tham Luang cave in Thailand's Chiang Rai province for more than two weeks left the hospital today where they were hospitalized. Police reported that the boys underwent a DNA test that showed that all were descendants of two parents, but there is no evidence that one or more of the boys did not remain among them while escaping the cave. The

Issue type: Grammatical error.

Original Text: Se dovesse darsi un voto, lei si darebbe un "sette e mezzo". Sette romani su dieci, però, secondo un sondaggio di Repubblica, non le darebbero nemmeno la sufficienza. Ad un anno e mezzo dal suo insediamento, è la dimostrazione più esplicita del limpido disaccordo che la popolazione ha con la sua amministrazione. From the review of A. Kersalme's "Odysseus" on the site of French journal "La Revue de Belles Lettres" (98th issue, March 2019). From the review of M....

Translation: If she had to give herself a grade, she would give herself a "seven and a half." Seven out of ten Romans, however, would not even give her a sufficiency, according to a Repubblica poll. A year and a half after taking office, it is the most explicit demonstration of the limpido disagreement the population has with her administration. From the review of A. Kersalme's "Odysseus" on the site of French journal "La Revue de Belles Lettres" (98th issue, March 2019). From the review of M....

Issue type: Factual error.

Generated Text: Passerà in carcere da un minimo di 42 anni fino all'ergastolo. Si è chiuso con la lettura di una sentenza durissima da parte del giudice della Superior Court di Los Angeles. La condanna dura fino al 2019, ma la rischia se si insulta o rinnega ancora una volta le sue azioni. Il procuratore generale della California, Kamala Harris, ha anche chiesto per lui la revoca di libertà condizionata. Il giovane è stato arrestato a Hong Kong nel maggio 2014....

Translation: He will spend from a minimum of 42 years to life in prison. It ended with the reading of a very harsh sentence by the Los Angeles Superior Court judge. The sentence lasts until 2019, but he risks it if he insults or disavows his actions again. California Attorney General Kamala Harris also sought revocation of parole for him. The young man was arrested in Hong Kong in May 2014...

Issue type: Grammatically correct, but does not sound natural in Italian and shows factual errors.

Table 6: Examples of mistakes made by pretrained Llama 7B. The smaller Llama 7b without fine-tuning shows several kinds of weaknesses generating both factual to grammatical errors. Out of 25 samples manually examined for this model, we find 0 contradicting the prompt, 5 switching to English, 8 unnatural Italian expressions and 6 grammatical errors.

Llama 7B Fine-Tuned

Generated Text: Passa per sette punti il programma di Matteo Salvini per bloccare I flussi e dare maggiore sicurezza al Paese. Il blocco dei porti a navi che battono bandiera di uno stato membro dell'Unione europea e non il Regno Unito, come annunciato da Salvini, sarebbe contro il diritto e il trattato, secondo l'ex ministro della Giustizia Franco Frattini. "Vi è un'interpretazione di diritto internazionale. A me pare che sia totalmente sbagliata", ha detto il politico liberale di Palazzo Chigi. "Sotto..."

Translation: It passes for seven points in Matteo Salvini's program to stop the flows and give more security to the country. Blocking ports to ships flying the flag of a European Union member state and not the United Kingdom, as announced by Salvini, would be against law and treaty, according to former Justice Minister Franco Frattini. "There is an interpretation of international law. It seems to me that it is totally wrong," said the liberal politician from Palazzo Chigi. "Under..."

Issue type: Factual errors.

Original Text: Alla vigilia dell'incontro, a Bruxelles, tra Jean-Claude Juncker e il presidente del Consiglio Giuseppe Conte, che sarà accompagnato dal ministro dell'Economia Tria, il commissario Ue Pierre Moscovici usa toni concilianti per suggerire che l'Italia non dovrebbe fare la spola tra l'Europa e le banche a credito. "Sarà difficile", dice il francese, "se la Banca centrale germanica non pone il suo veto alla pesca in una sola mano per un paese come l'Italia che sta in cerca di finanziamenti". "La..."

Translation: On the eve of a meeting in Brussels between Jean-Claude Juncker and Prime Minister Giuseppe Conte, who will be accompanied by Economy Minister Tria, EU Commissioner Pierre Moscovici uses conciliatory tones to suggest that Italy should not shuttle between Europe and banks on credit. "It will be difficult," says the Frenchman, "if the German Central Bank does not veto single-handed fishing for a country like Italy that is looking for financing." "The..."

Issue type: Unnatural expression.

Generated Text: Forti temporali si sono abbattuti nella mattinata di oggi su Milano e sul nord-ovest della Lombardia, spazzando via il caldo degli ultimi giorni. Il nubifragio ha creato qualche rallentamento sulle strade. Poche colture e alcune piantine. Questi sono i principali danni dei fenomeni atmosferici che hanno colpito l'Italia. L'Italia centro-meridionale ha visto 304.363 colture danneggiate, quasi 15mila piantine e 1.532 migliaia di alberi, una pioggia netta di oltre 10mila mila chili sul versante del Po e una coltura danneggiata...

Translation: Strong thunderstorms hit Milan and northwestern Lombardy in the morning today, sweeping away the heat of the past few days. The cloudburst created some slowdowns on roads. Few crops and a few seedlings. These are the main damages of the weather phenomena that hit Italy. South-central Italy saw 304,363 crops damaged, nearly 15,000 seedlings and 1,532,000 trees, a net rainfall of more than 10,000,000 kilograms on the Po River side, and a damaged crop...

Issue type: Factual error.

Table 7: Examples of mistakes made by Llama 7B fine-tuned on Italian. The smaller Llama 7b after fine-tuning generates flowing text but the facts are anyway less accurate than for larger models. Out of 25 samples manually examined for this model, we find 5 contradicting the prompt, 0 switching to English, 5 unnatural Italian expressions and 6 grammatical errors.

Llama 65B Pre-Trained

Generated Text: In centro Italia continuano le scosse e il numero dei morti sale. Paesi a foni e distrutti, palazzi accartocciati, mozziconi di chiese, rovine sinistre che fissano le piazze. E I corpi stanno fuori dalle case, senza casa, senza neanche la tomba che li ricuopre. Le indagini del Giornalista Enrico Lucci e le testimonianze dei superstiti. The BBC's Gavin Lee reports from Norcia, the central Italian town hardest hit by the quake. Italian Prime Minister Matteo Renzi has visited one of...

Translation: In central Italy, the tremors continue and the death toll rises. Aphonious and destroyed villages, crumpled buildings, church butts, sinister ruins staring into squares. And The bodies stand outside homes, homeless, without even the grave to cover them. Journalist Enrico Lucci's investigation and survivor testimony. The BBC's Gavin Lee reports from Norcia, the central Italian town hardest hit by the quake. Italian Prime Minister Matteo Renzi has visited one of...

Issue type: Grammatically correct, but does not sound natural in Italian.

Original Text: Due sbarchi in sole ventiquattr'ore a Lampedusa. Sono 84 gli immigrati giunti nell'isola tra venerdì e sabato e ospitati all'Hotspot di contrada Imbriacola. Tra loro ci sono anche I tre naufraghi di Tortuga, recuperati da un mercantile libanese. Non è ancora stata resa nota l'identità nazionale dei naufraghi, ma non mancano inchieste di persone che si sono rivolte ai carabinieri per cercare I loro cari. Lampedusa, 38 migrants - mostly from Tunisia - arrived safely on the island in...

Translation: Two landings in just twenty-four hours on Lampedusa. A total of 84 immigrants arrived on the island between Friday and Saturday and were housed at the contrada Imbriacola Hotspot. Among them are also The three shipwrecked Tortuga, recovered by a Lebanese merchant ship. The national identity of the castaways has not yet been released, but there is no shortage of inquiries from people who have turned to the Carabinieri to search for their loved ones. Lampedusa, 38 migrants - mostly from Tunisia - arrived safely on the island in...

Issue type: Switch to English

Generated Text: I numeri fanno spavento. Cinquemila disperati sbarcati solo nelle ultime quarantott'ore, quasi 65mila nei primi sei mesi dell'anno. Che l'operazione "Mare Nostrum", varata dopo la strage di Lampedusa, fosse riuscita a fermare il flusso di barche di migranti, lo abbiamo rilevato già da pochi mesi. Che la spinta verso l'Italia del prossimo potesse crescere di conseguenza, lo sapevamo. Che si dovesse progettare una strategia efficace e responsabile, in grado di dire dove e come, giustamente, stava l'opportunità...

Translation: The numbers are frightening. Five thousand desperate people landed in the last forty-eight hours alone, nearly 65,000 in the first six months of the year. That operation "Mare Nostrum," launched after the Lampedusa massacre, had succeeded in stopping the flow of migrant boats, we noted a few months ago. That the push to Italy of the next could grow as a result, we knew. That an effective and responsible strategy had to be designed that could tell where and how, rightly, the opportunity lay...

Issue type: Contradicts the prompt.

Table 8: Examples of subtle mistakes made by pretrained Llama 65B. While the quality of the language is almost on par with the prompts, written by professional journalists (for these short examples), the content is largely made up although in a realistic way. Out of 25 samples manually examined for this model, we find 2 contradicting the prompt, 3 switching to English, 7 unnatural Italian expressions and 2 grammatical errors.

Llama 65B Fine-Tuned

Generated Text: Gabriele Gravina è stato eletto presidente della Figg. Ha ottenuto il 92,2% dei voti (al primo turno serviva il 75%). Si chiude la gestione commissariale di Roberto Fabbicini, durata 9 mesi. L'ultimo presidente eletto era Carlo Tavecchio nel 2014, che aveva ottenuto il 63%. Gravina: "Riconoscendomi orgogliosamente l'apostrophe di figlio della Sicilia e della Normandia che le mie origini e la mia vita mi fanno, ho iniziato la mia esperienza nel calcio. Un calcio che ho giocato da bambino,..."

Translation: Gabriele Gravina has been elected president of Figg. He has obtained 92.2% of the votes (on first round he needed 76%), This ends the Roberto Fabbicini's technical administration, which lasted 9 months. The last president to be elected was Carlo Tavecchio in 2014, he scored 63%. Gravina: "Proudly acknowledging the apostrophe of son of Sicily and Normandy that my origins and my life make me, I began my experience in soccer. A soccer I played as a child,..."

Issue type: Factual errors.

Original Text: "È superficiale dire che è risorta la Dc". Parola di Francesco Rutelli che, intervistato dal Corriere, ripercorre le tappe della vita della Margherita, partito in cui hanno militato sia lui che Bertinotti. Rutelli l'ha lasciato da qualche anno e adesso è presidente dei Liberali per l'Italia, la lista che recentemente ha lanciato al Senato. E Bertinotti è tornato alla Dc per partecipare alle primarie del partito di D'Alema e Bersani, l'unico che ha il voto dei sostenitori della...

Translation: "It is superficial to say that the DC has risen again." Word of Francesco Rutelli, who, interviewed by Corriere, traces the stages of the life of Margherita, a party in which both he and Bertinotti militated. Rutelli left it a few years ago and is now president of Liberals for Italy, the list he recently launched in the Senate. And Bertinotti returned to the DC to participate in the primaries of D'Alema and Bersani's party, the only one that has the supporter vote of the...

Issue type: Factual errors.

Generated Text: Dopo 19 giorni, grazie all'accordo trovato dall'Europa, I 49 migranti di Sea Watch e Sea Eye sono sbarcati nel porto maltese di La Valetta, dove sono stati trasportati a bordo di autobus. Il ministro Migranti e Strade del Popolo Gabriele Toccafondi nei giorni scorsi ha incontrato a Parigi il ministro degli Esteri francese Jean-Yves Le Drian, il ministro dell'interno Christophe Castaner, il ministro della giustizia Nicole Belloubet e il ministro dell'istruzione Jean-Michel Blanquer. Si tratta di una missione di...

Translation: After 19 days, thanks to the agreement found by Europe, The 49 migrants from Sea Watch and Sea Eye landed in the Maltese port of La Valetta, where they were transported aboard buses. Migrants and People's Roads Minister Gabriele Toccafondi in recent days met in Paris with French Foreign Minister Jean-Yves Le Drian, Interior Minister Christophe Castaner, Justice Minister Nicole Belloubet, and Education Minister Jean-Michel Blanquer. This is a mission of...

Issue type: Factual errors.

Table 9: Examples of subtle mistakes made by Llama 65B fine-tuned on Italian. While the quality of the language is almost on par with the prompts, written by professional journalists (for these short examples), the content is largely made up although in a realistic way. Out of 25 samples manually examined for this model, we find 0 contradicting the prompt, 0 switching to English, 1 unnatural Italian expressions and 4 grammatical errors.